DOCUMENT RESUME

ED 129 910 95 TM 005 754

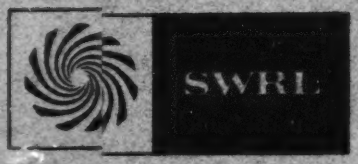
AUTHOR	Besel, Ronald
TITLE	Using Group Performance to Interpret Individual Responses to Criterion-Referenced Tests. Professional Paper 25.
INSTITUTION	Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	SWRL-PP-25
PUB DATE	25 Jun 73
NOTE	13p.
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	Algorithms; Correlation; *Criterion Referenced Tests; Decision Making; *Groups; Individual Differences; Mathematical Models; Norms; Performance; Probability; *Response Style (Tests); *Test Construction; *Test Interpretation
IDENTIFIERS	Mastery Learning Test Model

ABSTRACT

The contention is made that group performance data are useful in the construction and interpretation of criterion-referenced tests. The Mastery Learning Test Model, which was developed for analyzing criterion-referenced test data, is described. An estimate of the proportion of students in an instructional group having achieved the referent objectives is usable as a prior probability in interpreting individual responses. Considering instructional group performance enhances estimates of individual performance. Correlational data from a set of test items and a representative population of students are used to estimate the required item parameters. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED129910



SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 754

Using Group Performance to Interpret Individual Responses to Criterion-Referenced Tests

25 June 1973 Professional Paper 25

Published by SWRL Educational Research and Development, a public agency supported as a regional educational laboratory by funds from the National Institute of Education (NIE), Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position of NIE, and no official endorsement by NIE should be inferred.

SWRL EDUCATIONAL RESEARCH AND DEVELOPMENT

Professional Paper 25

June 1973

USING GROUP PERFORMANCE TO INTERPRET INDIVIDUAL RESPONSES TO CRITERION-REFERENCED TESTS

Ronald Bosel

ABSTRACT

The contention is made that group performance data are useful in the construction and interpretation of criterion-referenced tests. The Mastery Learning Test Model, which was developed for analyzing criterion-referenced test data, is described. An estimate of the proportion of students in an instructional group having achieved the referent objective is usable as a prior probability in interpreting individual responses. Considering instructional group performance enhances estimates of individual performance. Correlational data from a set of test items and a representative population of students are used to estimate the required item parameters.

USING GROUP PERFORMANCE TO INTERPRET INDIVIDUAL RESPONSES TO CRITERION-REFERENCED TESTS*

Ronald Besel

The proper use of norm-group data, both for the construction and the application of criterion-referenced tests, is an issue needing resolution. Typically "criterion-referenced" is defined in relation to norm-referenced (or standardized) tests. Livingston (1972) states that "norm-referenced measures compare the student's performance with the mean of a norm group whereas criterion-referenced measures compare his performance with a specified criterion score." On the basis of such definitions, Airasian and Madaus (1972) conclude that "the interpretation of a student's performance in a criterion-referenced situation is absolute and axiomatic, not dependent upon how other learners perform." Block (1971) observes that criterion-referenced "measurements are absolute indices designed to indicate what the pupil has or has not learned from a given instructional segment. The measurements are absolute in that they are interpretable solely vis-a-vis a fixed performance standard to criterion and need not be interpreted relative to other measurements."

These statements do not clarify the legitimacy or the value of norms in interpreting individual performance; they have led some to question the appropriateness of using any item selection procedure based on norm-group responses. It is contended here that norm-group performance is useful and legitimate information for both the construction and application of criterion-referenced tests.

A criterion-referenced test is here defined as a set of items sampled from a domain which has been judged to be an adequate representation of an instructional objective. This definition does not limit criterion-referenced tests to narrowly defined behavioral objectives for which an item form (Osburn, 1968) specifies how to generate every item in the domain. But, it is desirable that the domain be described in operational terms; using this description another test developer should be able to generate an equivalent domain of test items. The assumptions or theory relating the domain of items to the referent objective should be explicitly stated.

Procedures for selecting a sample of items from a domain depend upon the intended application of the test. One application of a criterion-referenced test is to estimate the proficiency of individual students relative to some achievement continuum (Kriewall, 1972). This appears to have been Glaser's (1963) original conception of the purpose of a criterion-referenced test, where he assumed that,

---

*A version of this paper was presented at the meeting of the American Educational Research Association, New Orleans, February, 1973.

"Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance." For applications where hand scoring of tests is used, a random or stratified random sampling of items from the domain permits the unweighted number of correct responses to be interpreted as a degree of proficiency measure. If computer scoring is used, a sample of highly discriminating items will yield a better estimate of proficiency. Thus, the rejection of sampling based on item discrimination indices (norm-group performance) is based on the assumptions that a degree-of-proficiency measure is required and that the test must be hand scored.

A frequent application of criterion-referenced tests is the making of categorical mastery, non-mastery decisions for students comprising an instructional group. Subsequent instruction for a student is contingent upon the category in which he is placed. Typically, test developers have computed a degree-of-proficiency index and then, on most frequently an arbitrary basis, selected a critical "passing" score. A problem that arises is that it is difficult, perhaps impossible, to define a meaningful degree-of-proficiency index for many types of legitimate instructional objectives. Ebel (1971) concludes that "criterion-referenced measurement may be practical in those few areas of achievement which focus on cultivation of a high degree of skill in the exercise of a limited number of abilities." Ebel's conclusion is based on the premise that a degree-of-proficiency scale "anchored at the extremities--a score at the top of the scale indicating complete or perfect mastery of some defined abilities; one at the bottom indicating complete absence of those abilities" is required. Fortunately, such a measurement scale is not needed for the categorical decision application.

THE MASTERY LEARNING TEST MODEL

The Mastery Learning Test Model has been designed to provide an appropriate algorithm for analyzing criterion-referenced test data for making the following instruction decision: "which students have achieved the referent objective." Two statistics are computed: the probability that a given student has achieved the objective and the proportion of an instructional group that has achieved the objective. The model assumes that each student in an instructional group can be treated as belonging to one of two groups--a group that has achieved the objective or one that has failed to achieve it. The two-state assumption does not deny the possibility of partial achievement of the objective. It does imply that categorization of students into two groups, masters and non-masters, is the desired type of decision and the basis for subsequent instruction.

The Mastery Learning Test Model and the true score theory upon which it is based are derived in an earlier paper (Besel, 1972).

This model is related to a simpler mastery testing model suggested by Emrick (1971). Emrick's model assumes that measurement error can be accounted for by two test parameters: α--the probability that a non-master will give a correct answer to an item; and β--the probability that a master will give an incorrect answer to an item. His model implicitly assumes that all item difficulties and inter-item correlations are equal. This assumption can be avoided by increasing the number of test parameters--either by permitting item α parameters, or item β parameters, or both.

PROBABILITY OF MASTERY ESTIMATION

Let,

x_{ij} represent the response of individual j to item i,

$$x_{ij} = \begin{cases} 1 \text{ if a correct response is given} \\ 0 \text{ if an incorrect response is given} \end{cases} \quad (1)$$

α_i = the probability that an individual in the $\bar{M}$ state will give a correct response to the i^{th} item.

β_i = the probability that an individual in the M state will give an incorrect response to the i^{th} item.

Using X to represent a response vector for a K-item test, Bayes formula can be used to estimate the conditional probability of mastery.

$$P(M/X) = \frac{PRM \cdot \prod_{i=1}^{K} P(x_i/M)}{PRM \cdot \prod_{i=1}^{K} P(x_i/M) + [1-PRM] \cdot \prod_{i=1}^{K} P(x_i/\bar{M})}, \quad (2)$$

where PRM is the prior probability of the mastery state. The j subscript was deleted to simplify notation. The denominator of equation (2) represents the prior probability of the response vector X.

ESTIMATING THE PROPORTION OF STUDENTS IN THE MASTERY STATE

Let,

$E(x_i)$ represent the expected value of x_i for a sample population of N students.

For an item with parameters (α_i, β_i),

$$E(x_i) \Big|_{M} = 1-\beta_i \quad \text{for the } N_m \text{ individuals in the mastery state.} \tag{3}$$

$$E(x_i) \Big|_{\bar{M}} = \alpha_i \quad \text{for the } (N-N_m) \text{ individuals in the non-mastery state.} \tag{4}$$

Then,

$$E(x_i) = \frac{1}{N} \left[N_m \cdot (1-\beta_i) + (N-N_m) \cdot \alpha_i\right] \tag{5}$$

Define proportion in mastery to be:

$$MP = \frac{N_m}{N} \tag{6}$$

An unbiased estimate of $E(x_i)$ is the proportion of students (PC_i) in the sample which gave a correct response to the i^{th} item.

Let GMP symbolize an estimate of the proportion in mastery, MP.

Then,

$$PC_i = GMP \cdot (1-\beta_i) + (1-GMP) \cdot \alpha_i \tag{7}$$

Solving for GMP yields

$$GMP = \frac{PC_i - \alpha_i}{1-\alpha_i-\beta_i}, \tag{8}$$

Since each item was assumed to be a measure of the same objective, the proportion in mastery, MP, for each item--or for a K-item test--must be equivalent. The GMP estimate for a K-item test can be shown to be

$$GMP = \frac{U/K - \bar{\alpha}}{1-\bar{\alpha}-\bar{\beta}}, \tag{9}$$

where,

$U = \sum_{i=1}^{K} PC_i$ is the test mean score;

$\bar{\alpha}$ is the average of the α_i;

$\bar{\beta}$ is the average of the β_i.

PRIOR PROBABILITIES BASED ON COLLATERAL INFORMATION

If mastery decisions are based upon responses to a small set of items sampled from a domain, it is likely that many errors of classification will be made. One way of obtaining more information on each examinee without requiring the administration of additional test items is to use the collateral information contained in the test data of other students (Hambleton and Novick, 1973). The proportion in mastery estimate computed using equation (9) can be used as a prior probability estimate. Group-based priors may increase accuracy to an extent equivalent to adding between 6 and 25 items to a test as short as 5 items (Novick, Lewis, and Jackson, 1973). While the use of group-estimated priors is somewhat controversial for selection decisions across instructional groups (Novick, 1970), it promises to enhance instructional decisions within an instructional group.

The probability-of-mastery measure is ideally suited for a decision-theoretic approach to selecting a cutting score for the mastery decision application. If L_1 and L_2 are used to represent the losses associated with false-fail and false-pass misclassifications, the appropriate cutting score on the probability-of-mastery measure can be shown to be:

$$P(M/X)_c = \frac{L_2}{L_1 + L_2} \qquad (10)$$

Only the ratio of L_2 to L_1 need be specified to derive a cutting score. If proportion in mastery is used as prior probability, the cutting score will decrease as the proportion in mastery estimate increases.

PARAMETER ESTIMATION

Both α and β item parameters can be estimated from the item response data collected from a representative sample of students. Two parameter estimation algorithms have been developed for a Mastery Learning Model which has a single test--β parameter and item--α parameters.* Least-squares estimates of the parameters are computed

*Computer program listings are available from the author upon request.

using three classes of empirical data:

1. Item difficulties

2. Inter-item covariances

3. Score histograms

The first algorithm computes the least-squares estimates using an independent estimate of the proportion of students that have achieved the referent objective (GMP). The second algorithm requires no input estimate of GMP: it is estimated from the data in addition to the α and β parameters.

The stability of the parameter estimates was evaluated, for each algorithm, using test data from the end-of-unit criterion exercises of the SWRL Beginning Reading Program. Data from two consecutive years (1970-71 and 1971-72) were sampled from schools participating in the quality assurance tryout of the SWRL reading program. Each criterion exercise measured the achievement of four program objectives: (1) words in a storybook, (2) word elements, (3) word attack (novel words), and (4) letter names. Five, three-option multiple-choice items were used for each objective. Data from all 10 units of the program were analyzed; the sample sizes shrank from 263 to 98 for the first year and from 418 to 173 for the second year.

The means and variances of the differences between the parameter estimates for the two years were examined (see Table 1). Computations were made for item α, average α ($\bar{\alpha}$), and test β. For the "Fixed GMP" algorithm two estimates of GMP were used. The first estimate was the proportion of students scoring 80% (4 right out of 5) or better for the outcome. The second estimate was the proportion with a perfect score. The item α differences are based on 50 items, average α and test β on 10 tests. The mean differences could be due partially to systematic differences in the student populations since different school districts were represented in the two samples. The variances are more appropriate estimates of parameter stability.

For the second algorithm (GMP not fixed) the variances vary considerably across outcomes. The "fixed-GMP" algorithm achieved uniformly better stability with the perfect score criterion noticeably better than with the 80% criterion. The variances for both item α and average α decreased as the difficulty of the objective increased. Letter names was the easiest objective, word attack the most difficult. The variances of test β, on the other hand, increased as the difficulty of the objective increased. This trend was apparent in all three sets of calculations for both algorithms. This result is consistent with the notion that ideally one would like to estimate β from the responses of a group--all of which have achieved the objective. Likewise the item alphas could be "best" estimated from a group--none of which have

Table 1. Stability of Mastery Learning Parameters
(Mean Difference/Variances of Difference)

Outcome	Parameter	Minimum Sum of Squares Solution	80% Criterion Solution	100% Criterion Solution
1 Storybook Words	Item α	-.081 / .0276	-.026 / .0191	-.013 / .0076
	$\bar{\alpha}$	-.081 / .0122	-.026 / .0031	-.013 / .0008
	β	.018 / .0006	-.002 / .0002	-.004 / .0001
2 Program Word Elements	Item α	-.059 / .0126	-.042 / .0170	-.041 / .0072
	$\bar{\alpha}$	-.059 / .0033	-.042 / .0015	-.041 / .0004
	β	-.003 / .0004	-.007 / .0005	-.006 / .0001
3 Word Attack	Item α	-.037 / .0083	-.032 / .0096	-.020 / .0043
	$\bar{\alpha}$	-.037 / .0011	-.032 / .0017	-.020 / .0007
	β	-.000 / .0006	-.001 / .0006	-.003 / .0001
4 Letter Names	Item α	..052 / .0956	-.026 / .0354	-.036 / .0080
	$\bar{\alpha}$	.052 / .0418	-.026 / .0101	-.036 / .0010
	β	-.004 / .0002	-.006 / .0002	-.004 / .0000

achieved the objective. When a mixed group is used, β is estimated most accurately when a high proportion of the group has achieved the objective. Lowering the GMP of the norm group improves the accuracy of the α estimates at the expense of β accuracy.

ITEM SELECTION

If the α and β parameters are estimated for a large sample of items from a domain, using an appropriate norm group, a small set of highly discriminating items can then be selected for future mastery-decision applications. The most promising item discrimination index is

$$\gamma_i = 1 - \alpha_i - \beta_i \qquad (11)$$

Items with a high γ index provide the most information for the mastery-decision application.

SUMMARY

The usage of an independent estimate of the proportion of students in a norm group which have achieved an objective resulted in significantly improved stability of mastery learning parameters. This should result in increased validity of the Mastery Learning Test Model for making categorical mastery/non-mastery decisions. This test model can be used to make mastery decisions on the basis of very short tests. Using the proportion-in-mastery estimate for an instructional group as a prior-probability results in improved estimates of the probability that an individual student has achieved the objective. Norm-group data can also be used to select the best set of items from a domain for the mastery decision application.

REFERENCES

Airasian, P.W. and Madaus, G.F. "Criterion-Referenced Testing in the Classroom." Measurement in Education, 3 (May, 1972), 1-8.

Besel, R.R. "A Mastery Learning Test Model." Paper presented at the 1972 Annual Meeting of the American Educational Research Association, April, 1972.

Block, J.W. "Criterion-Referenced Measurements: Potential." School Review, 79 (1971), 289-298.

Ebel, R.L. "Criterion-Referenced Measurements: Limitations." School Review, 79 (1971), 282-288.

Emrick, J.A. "An Evaluation Model for Mastery Testing." Journal of Educational Measurement, 8 (Winter, 1971), 321-326.

Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." American Psychologist, 18 (1963). 519-521.

Hambleton, R.K. and Novick, M.R. "Toward an Integration of Theory and Method for Criterion Referenced Tests." Paper presented at the 1973 Annual Meeting of the American Educational Research Association, February, 1973.

Kriewall, T.E. "Aspects and Applications of Criterion-Referenced Tests." Paper presented at the 1972 Annual Meeting of the American Educational Research Association, April, 1972.

Livingston, S.A. "Criterion-Referenced Applications of Classical Test Theory." Journal of Educational Measurement, 9 (Spring, 1972), 13-26.

Novick, M.R. "Bayesian Considerations in Educational Information Systems." In Proceedings of the 1970 Invitational Conference on Testing Problems. Educational Testing Service, October, 1970.

Novick, M.R., Lewis, C. and Jackson, P.H. "The Estimation of Proportions in 'M' groups." Psychometrika, 38 (1973), 19-46

Osburn, H.G. "Item Sampling for Achievement Testing." Educational and Psychological Measurement, 28 (1968), 95-104.